Institute for Scientific Computing Research

University Collaborative Research Program Subcontract Research Summaries

Multiscale Simulation Combining Direct Simulation Monte Carlo and Navier–Stokes Solvers

Berni J. Alder, PI
UC Davis

Yihao Zheng, RA
UC Davis

Alejandro L. Garcia, co-PI
San Jose State University

Andrew Wissink
LLNL, Collaborator

Summary

Numerical modeling of fluids is challenging when the physics of interest spans length scales differing by orders of magnitude. In the computation of hydrodynamic flows, structured adaptive mesh refinement (AMR) is used to efficiently perform calculations whose length scales span several orders of magnitude. However, when the finest mesh approaches the molecular scale, the macroscopic partial differential equations of fluid mechanics themselves are no longer valid. Adaptive mesh and algorithm refinement (AMAR) has been developed to solve these types of multiscale/multiphysics problems. In an AMAR calculation, at the particle scale a molecular algorithm such as the direct simulation Monte Carlo method (DSMC) is used. The objective of the current period is to develop for the continuum method a Navier–Stokes solver for integration with the DSMC method in the AMAR context.

An AMAR code has been developed at LLNL using Structured Adaptive Mesh Refinement Application Infrastructure (SAMRAI). The present code solves the inviscid Euler equations of gas dynamics at multiple (macroscopic) levels of refinement with the DSMC particle method.

We have replaced an existing SAMRAI inviscid Euler solver with a viscid Navier–Stokes solver and integrated the new solver into the existing SAMRAI scheme. The Navier–Stokes solver is a second-order Godunov-type explicit solver and has to date been tested under full multiregion, multilevel conditions for the following cases:

- one-dimensional acceleration-driven Poiseuille flow and
- two-dimensional pressure-driven Poiseuille flow.

The results agree well with a comparably refined stand-alone one-region one-level Navier–Stokes solver.

Our future plan is to apply the hybrid code combining the DSMC and the SAMRAI Navier–Stokes solver to study various classical hydrodynamic instability problems, such as flow in a pipe, with this refined computational tool.

Scalable Algebraic Domain Decomposition Preconditioners

Randolph E. Bank, PI
UC San Diego

Kathy Lu, RA
UC San Diego

Charles H. Tong and Panayot S. Vassilevski
LLNL Collaborators

Summary

We study the domain decomposition approach to the parallel solutions of partial differential equations with the novel feature that the subproblem residing on each processor is defined over the entire domain, although coarsely refined outside of the processor associated with the subdomain. This feature ensures that a global coarse description of the problem is contained within each of the subproblems. The advantages of this approach are that interprocessor communications are minimized while optimal order of convergence rates is preserved, and the convergence rate of the local subdomain solves can be optimized using the best existing sequential algebraic solvers.

This procedure is similar in philosophy to the parallel adaptive mesh refinement paradigm introduced by Bank and Holst, except that the present project deals with an algebraic version of the Bank-Holst paradigm in the sense that, instead of mesh refinement here we coarsen the degrees of freedom (or matrix) outside the prescribed subdomain. Thus, the domain decomposition method applied in each processor involves the local subdomain plus a small coarse space defined on the whole domain outside the processor, and each solve utilizes only a single interprocessor communication to retrieve the global vector. This approach can be applied to general sparse matrices, although matrices arising from discretization of partial differential equations are the principal target.

Besides the original global matrix representing the problem on the fine grid, the algorithm introduces a set of rectangular matrices (prolongation operators) and a set of coarse matrices that are constructed and stored in sparse parallel matrix format. A unique feature in our coarsening algorithm is that, even though coarsening is performed uniformly on the entire domain in parallel, to each processor coarsening occurs exclusively on the parts of the global linear system associated with other processors, and the matrix and right hand side for the local subdomain are preserved in the local subproblem construction.

We have focused mainly on developing and implementing a specific algebraic coarsening procedure that plays a central role in our overall algorithm. The software package, called FocusDD, consists of a parallel preconditioning algorithm built on top of the *hypre* library, developed at LLNL. As a by product of the Focus solver, an enriched set of functions to handle the *hypre* ParCSRMatrix and CSRMatrix data classes have been written and debugged. We extensively reuse existing subroutines in *hypre* for coarsening. For solving local subproblems, we use the SuperLU direct solver library. A detailed description of FocusDD is now available in reference manual form.

Since we now have an overall working code, our attention has turned to improving specific parts of the algorithm. In particular, we have been studying the linear systems arising from 5-point finite difference discretization of the Poisson equation on a rectangular domain. For this problem, we have made a matlab implementation of our

Randolph E. Bank, PI
UC San Diego

Kathy Lu, RA
UC San Diego

Charles H. Tong and
Panayot S. Vassilevski
LLNL Collaborators

Summary (continued)

algorithm, substituting a non-algebraic procedure for computing the prolongation matrix. Since this a very structured situation, we are able to easily construct a simple, structured prolongation operator. In the matlab code, we allow only one level of coarsening. We ran some numerical experiments and verified that the rate of convergence was less than 0.1, independent of both the size of the problem and the number of subdomains. We do not know if this prolongation operator is optimal in any sense, but the main point of the exercise was to demonstrate "proof of concept" in a simple setting. The FocusDD solver presently does not exhibit such ideal behavior for this model. Thus, we are presently examining the parallel coarsening procedure to determine what details in our implementation are responsible, and hopefully improve the overall performance.

In the longer term, our goal is to implement a more sophisticated coarsening procedure. In the current implementation, parts of the matrix on each processor are either fine or uniformly coarse. We have concluded from our initial analysis that it is useful to have graded coarsening. In the geometric sense, it means that near the boundary of the fine region, the coarsening should be modest. As one moves further from the fine region, the coarsening can be more generous. Thus, a geometric reduction of number of unknowns with distance from the subdomain seems appropriate. The overall order of the locally coarsened matrix A could be unchanged; it is its structure that will be modified. Our initial approach to achieve graded coarsening involves two different strategies. First, we will utilize our present parallel coarsening strategy (and software) to achieve a modest uniform coarsening. Each processor will then independently compute further coarsening to achieve the appropriate grading. While the algorithms used to achieve this grading will be sequential on each processor, all processors can do this step independently in parallel. The computation will be organized such that existing sequential graph coarsening software can be utilized, and less inter-processor communication will be required compared to parallel AMG in *hypre* library.

Our ultimate goal is to fine-tune our parallel code and benchmark its performance on selected applications of interest to LLNL. Our target computing platforms are massively parallel machines, including those at both UCSD and LLNL.

Similarity Query Processing for Quadratic Queries and Relevance Feedback

B. S. Manjunath and
S. Chandrasekaran, co-PIs
UC Santa Barbara

Helena Tesic
RA, UC Santa Barbara

Imola K. Fodor
LLNL Collaborator

Summary

The main objective of this research is to investigate user interactions with large multimedia databases in the context of direct access to and mining of the data. This work supports development of scalable algorithms for the interactive exploration of large, complex, multi-dimensional scientific data. The exponential growth of such multi-media and multi-sensor data in consumer as well as scientific applications pose many interesting and task critical challenges. There are several inter-related issues in the management of such data, including feature extraction, similarity-based search, high dimensional indexing, scalability to large data sets, and personalizing search and retrieval.

As technology advances and more visual data are available, we need more effective systems to handle the image data processing and user interaction. The framework must efficiently summarize information contained in the image data; it must provide scalability with respect to the nature, size and dimension of a dataset; and it must offer simple representations of the results and relationships discovered in the dataset. In particular, the database support to carry on the learning process in large image databases is an important issue that has been largely ignored. Our research focuses on similarity search and indexing of a real image/video databases that use relevance feedback mechanisms to improve retrieval results. We also introduce an efficient scheme that preserves the important data characteristics for data summarization.

Adaptive Nearest Neighbor Search in Relevance Feedback. We introduce the problem of repetitive nearest neighbor search in relevance feedback and propose an efficient search scheme for high dimensional feature spaces. Relevance feedback learning is a popular scheme used in content-based image and video retrieval to support high-level concept queries. This work addresses those scenarios in which a similarity or distance matrix is updated during each iteration of the relevance feedback search and a new set of nearest neighbors are computed. This repetitive nearest neighbor computation in high dimensional feature spaces is expensive, particularly when the number of items in the data set is large. We propose a search algorithm that supports relevance feedback for the general quadratic distance metric. The scheme exploits correlations between two consecutive nearest neighbor sets, thus significantly reducing the overall search complexity.

Dimensionality Reduction in Gabor Texture Features. Texture has been recognized as an important visual primitive in image analysis. A widely used texture descriptor is that computed us-ing multiscale Gabor filters. The high dimensionality and computational complexity of this descriptor adversely affect the efficiency of content-based retrieval systems. Our work shows how the dimensionality and complexity of the descriptor can be significantly reduced, while retaining comparable performance. This gain is based on a claim that the absolute values of the filter outputs follow a Rayleigh distribution. Ex-perimental results show that the dimensionality can be reduced by almost 50%,

B. S. Manjunath and
S. Chandrasekaran, co-PIs
UC Santa Barbara

Helena Tesic
RA, UC Santa Barbara

Imola K. Fodor
LLNL Collaborator

Summary (continued)

with a tradeoff of less than 2% on the error rate. We also propose a new normalization method that improves similarity retrieval and reduces indexing overhead. Details are discussed.

Data summarization and indexing for efficient similarity retrieval. This work introduces a conceptual representation for complex spatial arrangements of image features in large multimedia datasets. A data structure, termed the Spatial Event Cube (SEC), is formed from the co-occurrence matrices of perceptually classified features with respect to specific spatial relationships. A visual thesaurus constructed using supervised and unsupervised learning techniques is used to label the image features. SECs can be used to not only visualize the dominant spatial arrangements of feature classes but also to discover non-obvious configurations. Given a large collection of images with similar content—over space or time— associations are found among the image regions that visually characterize the dataset. These associations subsequently aid in the discovery of connections between image structure and semantic events, such as spatial rules. The emphasis is to mine the image content using little or no domain knowledge. An inclusive association rule algorithm is proposed as an extension of traditional association rules for application to image datasets. The definitions of standard transaction and association mining rules are extended for image datasets. Experiments show the approach has a wide range of application in image and video datasets.

This research has been presented at CASC seminars and at the Scientific Data Mining workshop held at IPAM in 2002 and has been documented in numerous refereed papers.

Simulation of Compressible Reacting Flows

Sutanu Sarkar, PI
UC San Diego

David Garrido Lopez, RA
UC San Diego

Andrew W. Cook
LLNL Collaborator

Summary

A fundamental problem of interest to NIF capsule design is the behavior of a burn front when it propagates through a deuterium/tritium (DT) mix contaminated by inert shell material introduced by Rayleigh-Taylor instabilities. In order to accurately describe burn front propagation, we have developed a scheme that is high-order on arbitrarily nonuniform grids. Nonuniform stretched grids are needed because the burn front, being much thinner than other flow features, needs a much smaller grid step for resolution of the transport and reactive processes that determine the burn velocity. We have simulated two-dimensional burn propagations into a contaminated mix.

Simulations on a nonuniform grid in physical space are often performed by using a computational grid that is uniform, and incorporating the Jacobian of the transformation between physical and computational grids into the governing equations. We refer to such an approach as Method 1. Although simple in principle, Method 1 does not ensure that the formal accuracy of the derivative in the case of uniform grid step is maintained on a nonuniform grid. The actual accuracy can be as low as first-order depending on the grid stretching as well as the value of the local derivative. In turbulent combustion, where sharp changes of the field variables occur locally, it is of interest to use a numerical scheme that has a defined high order of accuracy of the spatial derivative independent of the solution. One such approach, called Method 2, is to write an expression for the spatial derivative that, by construction, has the required spatial accuracy on an arbitrary nonuniform grid. We have demonstrated the advantage of Method 2 by evaluating the derivatives of known smoothly differentiable functions under both methods. Versions of Method 2 that are second, fourth and sixth-order accurate on arbitrary grids show the expected order of accuracy as resolution increases until roundoff error associated with very small grid spacing begins to be important. On the other hand, applications of Method 1 with derivatives that are fourth and sixth-order accurate on a uniform grid do not maintain high-order accuracy on the nonuniform grid. Therefore, Method 2 is used for the evaluation of spatial derivatives in simulations of the burn front.

We simulate burn front propagation in a reactant mix contaminated by a multi-scale field of inert fluctuations. Both Arrhenius chemistry and DT chemistry have been explored. The flame zone is thicker in the latter case; however, the chief results regarding the burn propagation are similar. The unsteady, three-dimensional form of the compressible equations for a reactive mix-ture of fluids is solved. A nonuniform grid is used in the direction of flame propagation, with clustering of points in the burn region. Derivatives in this direction are computed using a 6th-order compact scheme valid for arbitrary nonuniform grids. The code is parallelized using MPI and runs on the unclassified IBM SP machine at LLNL using up to 896×1024 grid points.

Sutanu Sarkar, PI
UC San Diego

David Garrido Lopez, RA
UC San Diego

Andrew W. Cook
LLNL Collaborator

Summary (continued)

In the zones contaminated by inert, the temperature rise due to the burn energy release is smaller than average leading to a lower reaction rate and a local burn velocity that is smaller than average. Thus, the nonuniform inert introduced into the reactant mix by flow instabilities leads to distortions of the burn front. An additional mechanism for distorting the burn front that is operative even in an uncontaminated mix is the so-called Darrieus-Landau (DL) instability. Linear stability analysis gives the growth rate of infinitesimal perturbations of an initially planar burn front. The value of an associated with the basic hydrodynamic instability increases with wavenumber but, due to the stabilizing thermal effect associated with heat diffusion, the growth rate eventually decreases at large wavenumbers. Thus, there is a preferred wavelength of maximal growth of the DL instability. At any given time, an 'average' burn velocity based on the overall reactant consumption rate can be defined.

The research was presented in the annual meeting of the APS Division of Fluid Dynamics and at a seminar in ISCR.

Probabilistic Clustering of Dynamic Trajectories for Scientific Data Mining

Padhraic Smyth
PI, UC Irvine

Scott Gaffney, RA
UC Irvine

Chandrika Kamath
LLNL Collaborator

Summary

Data-driven exploration of massive spatio-temporal data sets is an area where there is particular need of new data mining and data analysis techniques. Analysis of spatio-temporal data is inherently challenging, yet most current research in data mining is focused on algorithms based on more traditional feature-vector data representations. In this project we have developed a set of flexible and robust algorithms as well as software for tracking and clustering time-trajectories of coherent structures in spatio-temporal grid data. These algorithms and software provide a basic set of data analysis tools for exploration and modeling of dynamic objects, in a manner analogous to the much more widely-available techniques for clustering of multivariate vector data (e.g., k-means, Gaussian mixtures, hierarchical clustering, etc).

In our second year we have focused on extending our earlier probabilistic mixture-based modeling of trajectories of objects to a more general scheme known as "random effects" models. We have described a variety of algorithms for "trajectory clustering," including methods based on polynomial curve models and splines. The use of random effects models (also known as Bayesian hierarchical models) is shown to increase the predictive power of these clustering algorithms, by allowing each trajectory to have its own parameters (influenced by a population prior), which in effect leads to clustering in parameter space.

Clustering is typically used as a tool for understanding and exploring large data sets. Most clustering algorithms operate on so-called feature vectors of fixed dimension. In contrast to this we address the problem of clustering sets of variable-length curve or trajectory data generated by groups of objects or individuals. The curves are sequences of observations measured over time (or some notion of time), functionally dependent on an independent variable or set of variables. Typically, the independent variable is time, but in general it can be any number of variables measured over the same interval as the observations. It is assumed that the trajectory data is generated by groups of objects or individuals such as humans, animals, organizations, natural phenomena, etc. Unlike fixed-dimensional feature vectors, trajectories have variable lengths and can be observed at different measurement intervals as well as contain missing observations.

This type of data is quite common due in part to large-scale data collection in the scientific community. Such data cannot be clustered with any of the standard vector-based clus-tering algorithms absent some *ad hoc* preprocessing procedure to reduce the data to a set of fixed-dimensional feature vectors. Our focus is to model trajectory data directly using a set of model-based curve-clustering algorithms for which we use the all-encompassing name "mixture of regressions" or "regression mixtures."

For this project we have developed and extended several different types of regression mixture models to deal with curve data sets. All our algorithms explicitly model the dependence of trajectories on the independent set and all use an Expec-tation Maximization (EM) procedure to perform the clustering.

Padhraic Smyth
PI, UC Irvine

Scott Gaffney, RA
UC Irvine

Chandrika Kamath
LLNL Collaborator

Summary (continued)

We have applied of this approach to the problem of tracking and clustering of extra-tropical cyclones (ETCs) in the North Atlantic. Understanding ETC trajectories is scientifically important for understanding both the short-term and long-term dynamics of atmospheric processes. Atmospheric scientists are interested in the spatio-temporal patterns of evolution of ETCs for a number of reasons. For example, it is not well-understood how long-term climate changes (such as global warming) may influence ETC frequency, strength, occurrence and spatial distribution. Similarly, changes in ETC patterns may provide clues of long-term changes in the climatic processes that drive ETCs. The links between ETCs and local weather phenomena are also of interest: clearly ETCs have significant influence on local precipitation, and in this context, a better understanding of their dynamics could provide better forecasting techniques both on local and seasonal time-scales. Furthermore. an explicit model of ETC evolution can serve as an intermediate link between global atmospheric phenomena (e.g., geopotential height patterns and regimes) and local “weather-related” phenomena such as precipitation, addressing the long-standing problem in atmospheric science known as “downscaling” (i.e., how to link and model global-scale arid local-scale phenomena in a coherent manner). The data sets that we are working with are widely used simulation data sets known as general circulation model data. We have demonstrated that our clustering methodology, when applied to ETC data, reveals useful and interpretable information about ETC patterns and behavior.

Several software products have been produced during this project. Initially, all our code was prototyped in Matlab and later moved to a more efficient language. In more recent work we have designed, implemented, and tested a general-purpose curve clustering toolkit that takes any type of curve data and produces the desired clustering. The code is written in C and C++ and is highly optimized and integrated with the C-LAPACK mathematical library.

We have made three visits to LLNL since the start of thisp roject, and the doctoral student on the project has presented two seminars on this work to LLNL personnel. Several conference and journal papers document new methodological and scientific results under this project.

Statistical Inference from Microarray Data

Mark van der Laan, PI
UC Berkeley

Annette Molinaro-Clark, RA
UC Berkeley

Dan Moore
LLNL Collaborator

Summary

The overall goal of this collaboration is to link gene microarrays to disease progression for eventual applications in treatment decisions and disease prevention. This goal includes the following components: to assess the reproducibility and reli-ability of numerous summary measures on gene microarrays; to determine reliable methods for classifying tumors based on their genetic profile; to link the classified tumors to covariates and clinical outcomes; to link the gene microarrays to covariates; and to link these gene microarravs and covariates to clinical outcomes.

This year was comprised of work on three projects: utilizing genetic profiles to classify tumors; model selection when linking covariates to clinical outcomes; and prediction of survival when linking covariates to clinical outcomes.

The first project included a dataset containing 41 renal tumors measured on CGH arrays of 90 BACs with three sub-types of cancer and two benign groupings. The goal of this study was to determine whether loci showing changes can be used to discriminate between the sub-types of cancer and between the benign and non-benign groupings. This dataset allowed us to explore: methods of simultaneous inference; prediction of sub-type by gain and loss of chromosomal region; identification of homogeneous groups of tumors by using hierarchical clustering routines with varying distant metrics; clustering BACs; and clustering tumors within clusters of BACS.

The second ongoing project involves model selection when linking gene microar-rays and covariates to clinical outcomes. In this project we are addressing how to choose bivariate and trivariate models given hundreds or thousands of independent predictors. We are using simulated data and a clinical dataset for this project. For both, we observe data on time to recurrence, follow-up time, and covariates. Due to the possibility of multiple outcomes of interest we are implementing counting processes and mul-tiplicative intensity models. A special case of this is a counting process that is equivalent to the Cox Proportional Hazards model.

Our approach entails bootstrapping the datasets m times and subsequently split-ting each bootstrap into a test and validation set. The training set is used to estimate the coefficients from the multiplicative intensity model. In the bivariate models this is done four times for each combination of covariates. Once the coefficients are attained, the test set is used to estimate the partial log-likelihood. The mean of the m partial log-likelihoods is calculated for each model within each combination of variables. In order to choose those combinations that are most related to the clinical outcomes we are presently investigating different criteria to use on the ranked means.

The third project involves prediction of survival, or other clinical outcomes, by hundreds or thousands of covariates. We are interested in nonparametric regression methods including classification and regression trees, CART, which are conducive both to

Mark van der Laan, PI
UC Berkeley

Annette Molinaro-Clark, RA
UC Berkeley

Dan Moore
LLNL Collaborator

Summary (continued)

censored and complete (i.e., uncensored) data. In the past, numerous ap-proaches to adapt CART specifically to censored survival data have been suggested. However, these methods are strictly restricted to censored data and do not propose performance assessments of predictors given a user supplied loss function. In this project we are building predictors with CART, functional in both the censored and uncensored setting, and assessing their performance.

The algorithm for CART relies on a split criterion for splitting a node into two nodes for continuous (uncensored) outcomes. We have modified this criterion with an inverse probability of treatment weighted (IPTW) estimator, which allows for informative censoring and a gain in efficiency. After building the candidate pre-dictor using CART and the modified split criterion on a training set, we evaluate the performance of this predictor by estimating its conditional risk with the IPTW estimator of the risk based on a validation set. Our goal is to select a predictor, which has a risk approximating the optimal risk, i.e., the truth.